Interactive Feature Space Extension for Multidimensional Data Projection

Daniel Pérez[b,*], Leishi Zhang[c], Matthias Schaefer[a], Tobias Schreck[a], Daniel Keim[a], Ignacio Díaz[b]

[a] *Data Analysis and Visualization Group, University of Konstanz, Germany*
[b] *Área de Ingeniería de Sistemas y Automática, University of Oviedo, Spain*
[c] *Interaction Design Centre, Middlesex University, United Kingdom*

Abstract

Projecting multi-dimensional data to a lower-dimensional visual display is a commonly used approach for identifying and analyzing patterns in data. Many dimensionality reduction techniques exist for generating visual embeddings, but it is often hard to avoid cluttered projections when the data is large in size and noisy. For many application users who are not machine learning experts, it is difficult to control the process in order to improve the "readability" of the projection and at the same time to understand their quality. In this paper, we propose a simple interactive feature transformation approach that allows the analyst to de-clutter the visualization by gradually transforming the original feature space based on existing class knowledge. By changing a single parameter, the user can easily decide the desired trade-off between structural preservation and the visual quality during the transforming process. The proposed approach integrates semi-interactive feature transformation techniques as well as a variety of quality measures to help analysts generate uncluttered projections and understand their quality.

Keywords: Feature transformation, Dimensionality reduction, Multidimensional data projection

*Corresponding author. Tel.: +34 985182543
Email address: dperez@isa.uniovi.es (Daniel Pérez)

1. Introduction

Projection-based Data Analysis (PDA) is a widely used visual analytics approach for identifying and analyzing patterns in Multi-Dimensional (MD) data. The idea is to map each object in the data as a point to a two or three-dimensional visual display in such a way that similar objects are close to each other and dissimilar ones are further apart. The result is represented in a scatterplot where structures and patterns can be analyzed effectively. The mapping is usually achieved by a Dimensionality Reduction (DR) technique that approximates the distance (similarity) between objects in the MD data space to the Lower-Dimensional (LD) projection space. Fig. 1 shows an example of such projection.

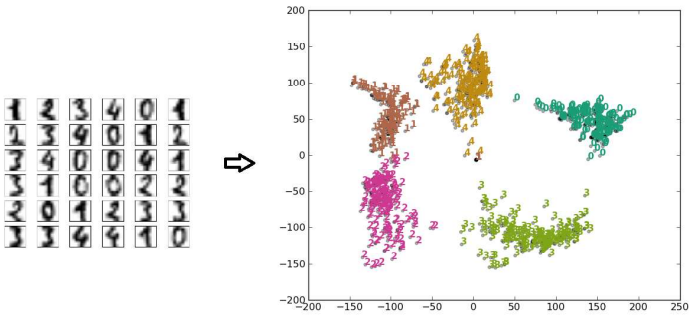

Figure 1: Images (28X28 D) of hand-written digits projected to a 2D display

A large number of DR methods exist [1, 2] for generating projections that preserve the original structure and characteristic of the data. However, when the data is large and noisy, the projection can be cluttered where points and groups overlap each other. The poor visual quality can make it difficult to identify and analyze patterns in the data. This problem originates from the *curse of dimensionality problem* [3]. First of all, distances measures tend to be less meaningful while dimensionality increases, as all objects become similar and dissimilar in many ways, leading to objects being plotted to similar locations in the visual display. Secondly when there is class information involved, those features that are irrelevant to the class labels can obscure the class separation, leading to blurred group boundaries in the projection.

For PDA it is important that the projection not only preserves the data structure but also reveals patterns in the data. When class information is available, a common approach is to take a supervised DR approach that

uses class labels to improve group separation in the projection. Available methods include the *Linear Discriminative Analysis* (LDA) [4] that extracts the discriminative features to the class labels and use them to generate embedding, the *Neighborhood Components Analysis* (NCA) [5] that learns a distance metric by finding a linear transformation of input data such that the average classification performance is maximized in the projection space, and the *Maximally Collapsing Metric Learning* (MCML) [6] that aims at learning a distance metric that tries to collapse all objects in the same class to a single point and push objects in other classes far away.

Supervised DR helps improve visual clarity of projections but an uncluttered projection can hardly be guaranteed. On the other hand for explorative analysis, it is important to gain an overview of the data before detailed analysis [7]. A recent work by Schaefer et al. [8] proposed a novel approach that improves the visual quality of the projection by adding class-related features to the original feature space and generating projections based on the extended data. Some promising results were reported. It is not surprising that by feature extension the original structure of the data will be distorted to a certain degree. However, the paper shows that a good compromise between the structural preservation and visual quality can often be made. Moreover, when the data is large and noisy, the method often distorts the structure in a good way such that meaningful patterns obscured by the noise can be revealed especially when the class labels fit to the data structure.

Another issue of PDA is interactivity and transparency. For many application users who are not Machine Learning (ML) experts, the DR process is often kept in a black-box which makes it difficult to understand and control. Recent advances in solving this problem include i) the interactive visual DR approaches that integrate the human expertise in the DR process [9, 10], ii) the interactive MD projection system that allows the user to manipulate the control points (subset of sample points) in the visual space based on their knowledge to better organise them as groups [11] and iii) the interactive feature space transformation approach that allows the analyst to transform existing feature space using different strategies based on their knowledge and understanding about the data [8, 12].

In this paper, we present a simple but effective interactive approach that allows the analyst to improve the visual quality of the projection by gradually transforming the original feature space towards clearer group separation in the projection space. This group separation helps in the exploration but it is restricted by the underlying data support. The approach is similar

to supervised DR but provides additional user control over the transformation process. The user can adjust the degree of transformation, via a single weighting parameter, and stop at any point where a projection is obtained. The method can be applied on top of any existing supervised DR approach to further improve the group separation. When class labels are not available, clustering results can be used as substitutions to support explorative analysis. In such case, more uncertainty is often introduced, however a series of quality measures are provided to help understand the quality of the projection both in terms of structural preservation and visual clarity. These quality measures provide additional numerical evaluation for the decision of a final projection.

The main contributions of this paper include 1) a novel and flexible visual analytics approach that combines interactive visualization, feature transformation, and quality evaluation for PDA; 2) a simple but effective feature transformation technique for gradually improving group separation in the projections space; 3) an interactive user interface that provides user control over the transformation process. The remainder of this paper is organized as follows. In Section 2 we discuss related work, in Section 3 we explain the details of the proposed approach, in Section 4 we demonstrate the effectiveness of the method with data by means of a set of experiment results, in Section 5 some characteristics and limitations of the method are discussed and finally, in Section 6 we draw conclusions with an outlook over future work.

2. Related work

The work presented in this paper relates to interactive MD data projection, feature transformation and quality assessment of visual embedding.

2.1. Interactive MD data projection and Feature Transformation

Classical DR methods estimate the structure of manifolds with a smaller intrinsic dimensionality. When used for generating visual embedding of MD data, the result can be unsatisfactory, especially when the dimensionality is high and the data contains noise. Firstly, the projection space is limited to 2D or 3D. Secondly, by its nature the reduction causes information loss and it is often difficult for the algorithms to determine which information is less relevant to the analysis tasks. In [13] the importance of integrating interactions with statistic methods (in particular, DR techniques) to support exploratory analysis of MD data is discussed. By interactive analysis, the analyst can

better steer the DR process by incorporating their domain knowledge and analytical skills for generating better projections.

In recent years, the idea of interactive projection has been widely adopted. For example, a semi-supervised approach is proposed in [14] for projecting MD data. In [15] interactive projection techniques are developed to allow the analyst integrate their knowledge about the data to the DR process. The *iPCA* [9] is proposed to provide coordinated views for interactive analysis of projections computed by PCA. In [10] the *iVisClassifier* system that integrates supervised DR technique LDA with interactivity is developed. The analysis of DR techniques with interactive controls were also proposed in [16] and the *DimStiller* framework [17] where the user is guided during the analysis process by means of workflows.

An effective approach to improve the visual quality of the projection is feature transformation. Given grouping information such as class labels or natural groups (clusters) in the data, the analysts may want to improve the visual quality of the projection gradually so that detailed analysis can be carried out. This can be achieved by pulling group members closer to each other in the projection and pushing non-group members further apart in the projection space. In theory such a task can be fulfilled by supervised DR, however, as discussed in the previous section the fully automatic approach lacks user control and transparency. Schaefer's approach [8] improves the existing solution by allowing the analyst to extend certain features in the data based on grouping information and to add the extended features to the original feature space for generating better quality projections. The result shows that a good compromise can often be made between structural preservation and visual clarification. In [11, 18] another user-driven feature transformation approach, the *Local Affine Multidimensional Projection (LAMP)* is proposed and implemented. *LAMP* allows the user to modify the point locations in the visual display and use the modification as feedback to update the original feature space in order to achieve better visual quality. The approach provides easy user control over the projection process and does not require much ML knowledge. However when the location of multiple points are modified in the visual display, the method may encounter heavy computation load while updating local neighborhood diagrams of multiple control points. Another interesting approach called *Dis-Function* was proposed by Brown et al. [19] which displays the projection on an interactive visual display such that the analyst can move points around to modify the distance between objects based on their own knowledge. The modification on

the visual space is then used to update the distance function and recompute the distance measure. Such approach integrates new knowledge to the data which is similar to our approach, except that Dis-Function requires some prior knowledge of distance between objects, and our approach is meant for using existing grouping information.

In addition to the above mentioned work, a comparison of feature sets can be found in [20], where an interactive exploration can be made for the selection of suitable data descriptors. A related problem was addressed in [21] where dendrogram structures were extracted from alternative feature sets, and applied for interactive comparison and selection of feature sets. These interactive methods demonstrate the possibility of improving PDA by incorporating user knowledge and feedback. However interactive MD data projection remains a challenge as many of the existing methods are either dependent on a particular DR technique, or rely on a good understanding of the applied DR techniques.

2.2. Quality Metrics

Despite the large number of DR techniques that have been developed, the question of quality assessment of a given projection has only been studied in several cases and systematized in recent years [22, 23].

The first measures introduced to assess the quality of a projection were the so called *stress* and *strain* measures [24, 25]. These measures assess the quality of structural preservation by computing the differences of the pairwise distances between objects in the LD embedding and the corresponding distances in high-dimensional (HD) data space. They come from objective functions of a family of DR techniques such as Multidimensional Scaling (MDS) so that errors can be evaluated at the end of the minimization of the function. For example, one of the most commonly used *Sammon's stress* refers to the final value of the error function in the Sammon's projection algorithm as follows,

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \tag{1}$$

where d^*_{ij} is the distance between two points i and j in the HD data space and d_{ij} is the distance between the corresponding points in the LD projection space.

While *strain* and *stress* measures analyze the preservation of data structure based on differences of distances, several measures like *trustworthiness* and *continuity* [26] and the *K-ary neighborhoods* measure [27] assess the quality of a projection in a broader applicability, taking into consideration also neighborhood preservation using rank-based criteria.

For example, the *K-ary* measure is defined as

$$Q_{\mathrm{NX}}(K) = \sum_{i=1}^{N} \frac{|n_i^K \cap v_i^K|}{KN} \ , \tag{2}$$

where n_i^K and v_i^K are the K nearest neighbors of the point i in HD and LD spaces respectively. It is usually displayed as a line for the different values of K from 0 to $N-1$, here the average of these values Q_{avg} is considered in order to summarize the overall quality in a number between 0 and 1, where the higher value indicates better projection. Beside, when the data is labelled, the classification error is a typical choice, see for instance [28] and other references in [29]. The integration of classification error measures in the DR technique leads to better group separation in the final embedding.

Apart from the structural preservation quality measures mentioned above, a set of visual quality measures has also been developed. Examples include *Histogram Density Measure* that ranks scatter plot visualizations of multidimensional data, the *Class Density Measure* that assess class separation of a given projection, both proposed in [30], and class consistency measures [31]. Moreover, the *overlap measures*, defined in [8], compute the overlap area between groups and overlap object density in a multidimensional data projection. The overlap area sums the area of all the overlap regions $intersect(i,j)$ between pairwise groups for the set g of groups:

$$ov_{reg} = \sum_{i=1}^{|g|-1} \sum_{j=i+1}^{|g|} intersect(i,j) \tag{3}$$

The overlap regions are computed from the definition of the region of each group described by using the concave hull of the objects of each group proposed in [32]. The overlap density takes into account the density of the points over-plotted in the visual display. The visual display is divided into grids units where the occupation of a specific class is determined by using Gaussian functions G in the function f defined as follows

$$f(G_{ip}, G_{jp}) = \begin{cases} 1 & if \quad G_{ip} > 0 \ and \ G_{jp} > 0 \\ 0 & else \end{cases} \tag{4}$$

Thus, f is activated by 1 in the case where a grid square unit is occupied by two classes with the Gaussian model. The sum of these grids found for pairwise classes gives the overlap density measure, defined for K classes and an image of P pixels as

$$ov_{density} = \sum_{i=1}^{|K|-1} \sum_{j=i+1}^{K} \sum_{p=1}^{P} f(G_{ip}, G_{jp}) \tag{5}$$

In the next examples, the grid resolution is uniformly set to 3 pixels and σ value of the Gaussian model to 12, so that different experiment results can be compared.

3. Interactive feature extension

In this paper, we propose an analysis framework that combines the transformation of the feature space, the interactive parameter setting and visualization to help analysts achieve a better interpretation of projection results. Given a MD dataset, available grouping information is used to generate an extended feature space in such a way that the class knowledge is introduced in the extended feature space. The analyst can select certain attributes or the whole set for feature extension based on their knowledge and modify the projection gradually in order to achieve a good visual embedding. The quality of the projections will be evaluated using various quality measures. The process can be repeated iteratively until a satisfactory projection is achieved. Fig. 2 shows the flowchart of the proposed method.

3.1. Feature space extension

The basic idea of the feature space transformation is to extend certain features based on available grouping information. Consider a MD dataset as a matrix $\mathbf{X}$ where rows are data items and columns are features, and the labels $\mathbf{y}$ are given to the class corresponding to the i-th row.

$$\mathbf{X} = [x_{ij}] \in \mathbb{R}^{n \times d} \qquad \mathbf{y} = [y_i] \in \mathbb{N}^n \tag{6}$$

With $i = 1, \ldots, n$ and $j = 1, \ldots, d$, where n is the number of feature vectors and d the number of dimensions. If m features are selected $f = f_1, \ldots, f_m$, the extended data matrix $\mathbf{X}'$ is defined as follows,

$$\mathbf{X}' = [x_{ij} \ |\tilde{x}_{ij}] \in \mathbb{R}^{n \times (d+m)} \tag{7}$$

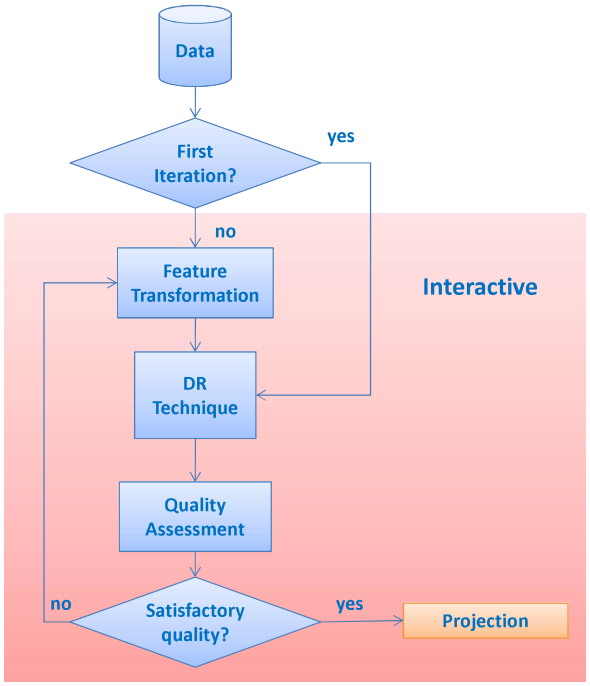

Figure 2: Workflow of the method

where $\tilde{x}_{ij}$ is the statistical value corresponding to the class label y_i in the feature f_j. Here, we use the arithmetic mean within the class members on a particular dimension. For example suppose we have a dataset with 2 attributes, 4 records that belong to 2 classes as shown below:

Sepal.Length	Sepal.Width	Species
5.1	3.5	*setosa*
4.9	3.0	*setosa*
7.0	3.2	*versicolor*
6.4	3.2	*versicolor*

The class labels are used to compute the mean values for each class in each dimension:

	Sepal.Length	Sepal.Width
$mean_{setosa}$	5.0	3.25
$mean_{versicolor}$	6.7	3.2

The $\mathbf{X}'$ matrix can be built as an extension of the original data $\mathbf{X}$ as following:

Original		*Extended*		
dim1	dim2	ext1	ext2	Species
5.1	3.5	5.0	3.25	*setosa*
4.9	3.0	5.0	3.25	*setosa*
7.0	3.2	6.7	3.2	*versicolor*
6.2	3.0	6.7	3.2	*versicolor*

Both original and extended space will be combined together to decide the distance metric for DR. Although we use class as an example statistical value for $\tilde{x}_{ij}$ in the above example, it should be noted that $\tilde{x}_{ij}$ can be many other statistical values such as median or other form of averages. An effective approach would be to decide which statistical values to use for each dimension based on the data distribution. Detailed discussion relating to this issue can be found in [8]. For all the experiments in this paper we extend mean values based on class labels for simple illustration purpose.

3.2. Weighted extension of the feature space

Having the data matrix $\mathbf{X}$ and labels $\mathbf{y}$ (see Equation 6), as explained above, the extended data matrix $\mathbf{X}'$ is defined by the original matrix $\mathbf{X}$ and the extended part $\tilde{\mathbf{X}}$ as follows:

$$\mathbf{X}' = \left[\mathbf{X} \mid \tilde{\mathbf{X}}\right] \tag{8}$$

Assuming the extension of the whole set of features and using mean values of each class labels, $\mathbf{X}' \in \mathbb{R}^{n \times 2d}$. In this case, $\tilde{\mathbf{X}}$ is composed by the centroids of the corresponding class described by the labels.

$$\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_i] \in \mathbb{R}^{n \times d} \quad \text{being} \quad \tilde{\mathbf{x}}_i = \frac{1}{|C_{y_i}|} \sum_{i \in C_{y_i}} x_{ij} \tag{9}$$

where C_{y_i} is the set of indices of samples belonging to class y_i.

A real parameter $\lambda \in [0, 1]$ allows the gradual transition between original data ($\mathbf{X}$) and the extended part ($\tilde{\mathbf{X}}$) by applying a simple change in the metrics of the extended feature space given by $\mathbf{X}_{weight} = \mathbf{X}'\mathbf{W}_\lambda$, being the matrix $\mathbf{W}_\lambda \in \mathbb{R}^{2d \times 2d}$ as follows:

$$\mathbf{W}_\lambda = \left(\begin{array}{c|c} (1-\lambda)\,\mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \lambda\,\mathbf{I} \end{array}\right), \ \lambda \in \mathbb{R} \tag{10}$$

Therefore, $\mathbf{X}_{weight}$ is the weighted data matrix used for computing low-dimensional embedding. The parameter λ can be changed interactively so that the user can trade between inter-class and intra-class topological organization of data. In this way, with $\lambda = 0$ the projection reveals the structure to the original dataset and with $\lambda = 1$ only to the applied extension. Thus, a good starting point for this analysis is the weighted extension of the whole feature space so that the analyst can change easily the embedding or return to the original. This is achieved only by interacting with λ parameter to obtain a more meaningful projection, assessed both visually and by quality measures.

Note that our proposed method is independent of the DR technique that computes the projection, hence it inherits the same level of computational complexity of the applied DR technique. However, various scalability approaches that involve sub-sampling and approximation have been made towards handling large data. For example, Li et al. [33] proposed a scalable scheme that improves the efficiency of the *Singular Value Decomposition* (SVD) process by first sampling a subset of columns from the input matrix and then approximate SVD on the inner sub-matrix using matrix approximation algorithms. Yang et al. [34] proposed an optimization approach that reduces the computational cost of *Neighbor Embedding* methods by computing close-by points individually but approximating far-away points by their center of mass. Bunte et al. [35] proposed a relevance learning approach that incorporates prior knowledge of the data such that the computational cost can be saved by reducing the number of adaptive parameters. Our proposed method can be used in conjunction with these methods to achieve better scalability.

3.3. An illustrative example

Here a very simple example is presented to illustrate the proposed method. The data consists of two Gaussian clusters of 150 points each with a small overlap in 2 dimensions. The method is applied to the data following the weighted extension of the feature space using the mean values for each class corresponding to each cluster. The DR technique to compute the projections is *PCA*. In Fig. 3 the resulting projections are represented for several values of the parameter λ. The projection for $\lambda = 0$ corresponds to the original data where the two clusters are not fully revealed. As the λ parameter increases, the projection changes revealing the grouping information. Since the distances inside of each cluster are not modified, the local structure in each

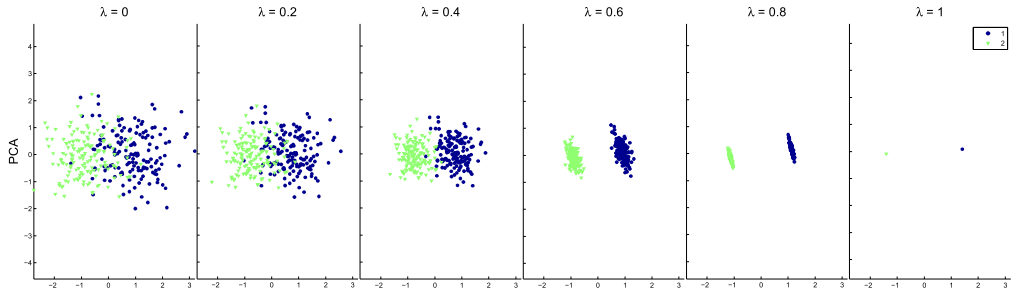

Figure 3: Projections from the proposed method applied to 2D clusters data example using PCA.

cluster is preserved for the new projections. For the highest value ($\lambda = 1$) the projection is purely based on the grouping information (mean value of each cluster), therefore all the data points are pushed to the centroids of corresponding clusters. Therefore this can be considered as a representation of the classes distribution. For a correct interpretation of the original data structure the projections for high values of λ are not useful and can be neglected. The interaction by means of the λ parameter provides control to the user and improves the exploration tasks. Moreover, a numerical evaluation of the projection gives more information to the user in order to judge the optimum point of the transformation, this is explained in more detail in Section 4.5.

4. Experiments and Results

In this section we evaluate the proposed method with different datasets and use cases. The datasets are selected representing data of various dimensionality, number of classes, synthetic and real (see Table 1). Four use cases are designed to test the method from different perspectives, including:

- c1: synthetic vs. real data - the first use case applies the method on two synthetic examples, the remaining use cases are applied on real data.
- c2: time series data - this use case shows an example of improving visual clarification of projections for analyzing patterns in time series data.
- c3: extending full feature space vs. selected features - the third use case demonstrates the potential of improving the effectiveness of the approach by extending only a subset of the features based on knowledge and understanding about the data.

Name	Size	Dimensions	Classes
3D clusters	500	3	5
synthetic-gaussian	500	10	5
eCons (Weekday)	338	24	7
eCons (Month)	338	24	12
hiv	78	159	6
yeast	1452	7	10

Table 1: Description of data

c4: supervised vs. unsupervised DR - the last use case applies the method on two supervised DR methods, while the other use cases apply unsupervised DR techniques.

All the experiments start with an original projection generated by a standard DR technique with λ value set to 0. For unsupervised DR we apply *PCA* and *t-SNE* that are widely used by the visualization community for explorative data analysis. For supervised DR we choose two recent advances including *NCA* and *MCML* as introduced in Section 1. The original feature space is extended using the weighted extension strategy as described in Section 3. All the projections were computed using Matlab implementations of DR algorithms from the toolbox [2] or the original authors. The projections of the original and extended feature space are computed using the same DR technique with the same parameter setting, after a z-score normalization. Where the t-SNE technique is applied and the new projection requires the perplexity parameter to be updated, we regenerate a new projection using the new parameter setting to replace the original projection for comparison. Since the performance of t-SNE is quite robust in terms of variation on perplexity values, such updates does not usually change the original projection to a great extent. Mean and standard deviation of the quality measures are computed after 10 iterations for this technique. In order to make more comparable projections, a linear transformation determined by *procrustes analysis* [36] is performed between projections.

Next we illustrate the results of the experiments. The evaluation of the projections of these experiments are discussed in Section 4.5.

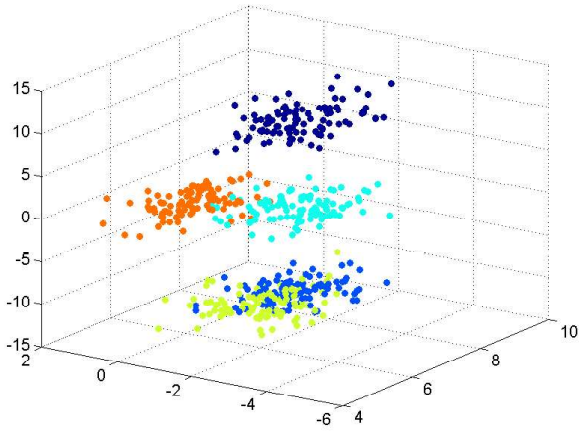

Figure 4: Representation of *3D clusters* data example.

4.1. Synthetic examples

4.1.1. 3D clusters example

To illustrate the idea conceptually, we apply the proposed method in a synthetic dataset that contains 3 dimensions. The data consists of 5 Gaussian clusters each containing 100 samples. Fig. 4 shows the original structure of the data in a three dimensional coordinate system. As shown in the figure, the top three clusters (colored in navy, cyan and orange) are well separated, but the two clusters at the bottom of the display (colored in blue and green) overlap against each other.

Based on the cluster labels a series of weighted feature extension were added to the original data, with the weight (λ parameter) set to several values. Fig. 5 shows two dimensional projections of the transformed data generated using PCA (upper) and t-SNE (lower) technique with a perplexity value of 20.

As as one can see from the figure. The process of transforming data by integrating group information modifies the original data space in such a way groups are better separated in the projection. In this particular example, when the values of λ is between 0.4 and 0.6, both DR techniques generate projections with clear group separation. Naturally when λ value is 0, the projection is purely based on the original data, hence in terms of the

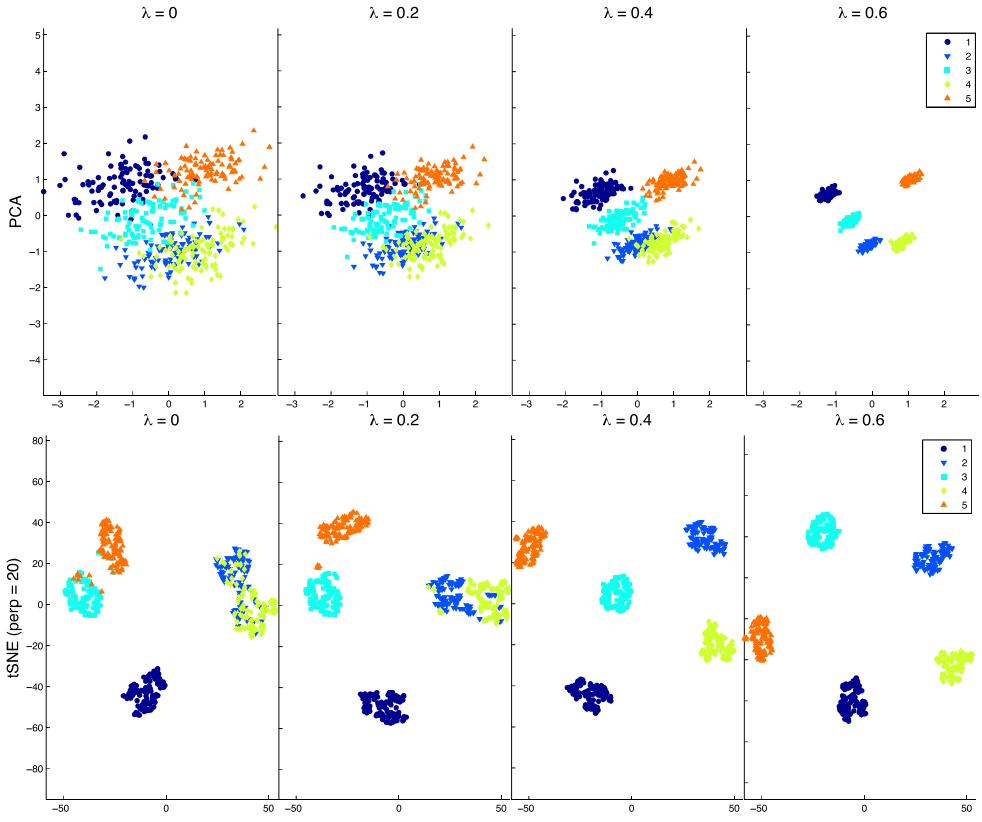

Figure 5: Projections of *3D clusters* with weighted extension for several λ values using cluster information, computed by PCA (top) and t-SNE (bottom) with a perplexity of 20.

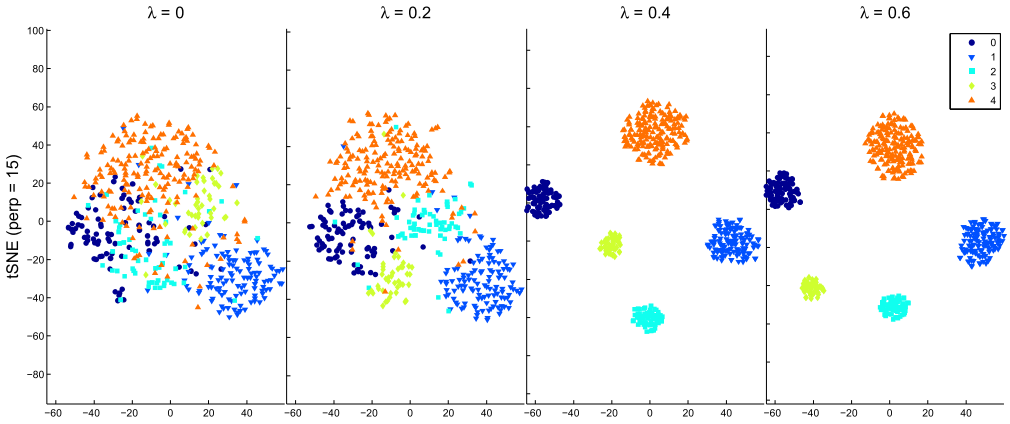

Figure 6: t-SNE projections of *synthetic-gaussian* dataset with weighted extension for several λ values using cluster information with a perplexity of 15.

preservation of the original data structure, there is no difference between the proposed method and the standard DR technique that is applied for generating the projection. Furthermore, it can be considered as an initial reference view to compare the new projections obtained by the method.

4.1.2. Synthetic gaussian example

In this experiment a synthetic dataset is used to evaluate the proposed method in a simple scenario. The data consists of 5 random Gaussian clusters of 10 dimensions and is generated from the work in [37]. Fig. 6 shows the resulting projections computed using t-SNE technique with a perplexity value of 15. Different colors and markers are used to distinguish the 5 different clusters. As one can see, compared to the projection of the original data, the projections of transformed data (with λ values between 0.2 and 0.4) provide a clearer separation of clusters. Even without color coding, it would not be difficult for the analyst to identify the patterns revealed in the projection.

Similar projections can be obtained using t-SNE for high values of λ. This is due to the fact that t-SNE aims at preserving both the local and global structure, where the importance of modeling the separations of datapoints is almost independent of the magnitudes of those separations [38].

4.2. Time series data

In the second use case we use the *eCons* data that records the active power usage at a university building over a year. The dataset is aggregated to

338 days (samples with missing values removed) and 24 attributes (value for each hour per day). The task is to identify different types of daily consumption patterns. In this experiment different types of temporal information, such as weekday or month, are incorporated into the resulting projection. The analysis of daily consumption patterns was addressed previously, in [39] where a calendar view combined with cluster information is used for an effective exploration of time series data. While the calendar view provides a good platform for univariate time series analysis, in this case our approach is designed more towards projecting multivariate data with the flexibility of adjusting time intervals and selection of class knowledge.

Two types of class labels are considered: classes corresponding to the type of the weekday (1-Sun; 2-Mon; ... 6-Fri; 7-Sat); and the corresponding month (1-January; ... 12-December), respectively.

First we analyze the data based on days of the week. The projections are computed using t-SNE with perplexity value set to 20. Different colors are assigned to different days of the week. In the projection of the original data for $\lambda = 0$, one can easily see two distinct groups (see Fig. 7, top). The result can be interpreted as "working days" and "non-working days". However in the embedding of extended data for $\lambda = 0.2$ (see Fig. 7, top), we see more interesting patterns, for example, most of the Mondays (blue triangle) appear to be in a separate cluster. Furthermore, the "non-working days" cluster splits into "weekends" and "bank holidays" clusters.

At the bottom of Fig. 7 an analogous analysis of the same dataset is shown. In this figure, colors are used to differentiate which months does a date belong to. The projections are again generated using t-SNE and the perplexity value is set to 20. The projection of the original data (left) shows two distinct groups (high- and low-consumption clusters) as in the previous case. Each group with a mixture of dates that belong to different months. However the projection of the extended data ($\lambda = 0.4$) further separates August dates from the rest of the points revealing a remarkable behaviour inside the low-consumption cluster. This could be explained by the university holiday period throughout August. In addition, months such as February, March and October show different consumption patterns from the rest.

Note that the new projections modify the location of the points from original clusters taking into account the information that the user incorporates. In this case, the same initial projection, showing two main clusters of daily consumption, is modified by two different criteria (weekday and month), that divide these groups revealing the introduced information hierarchically with

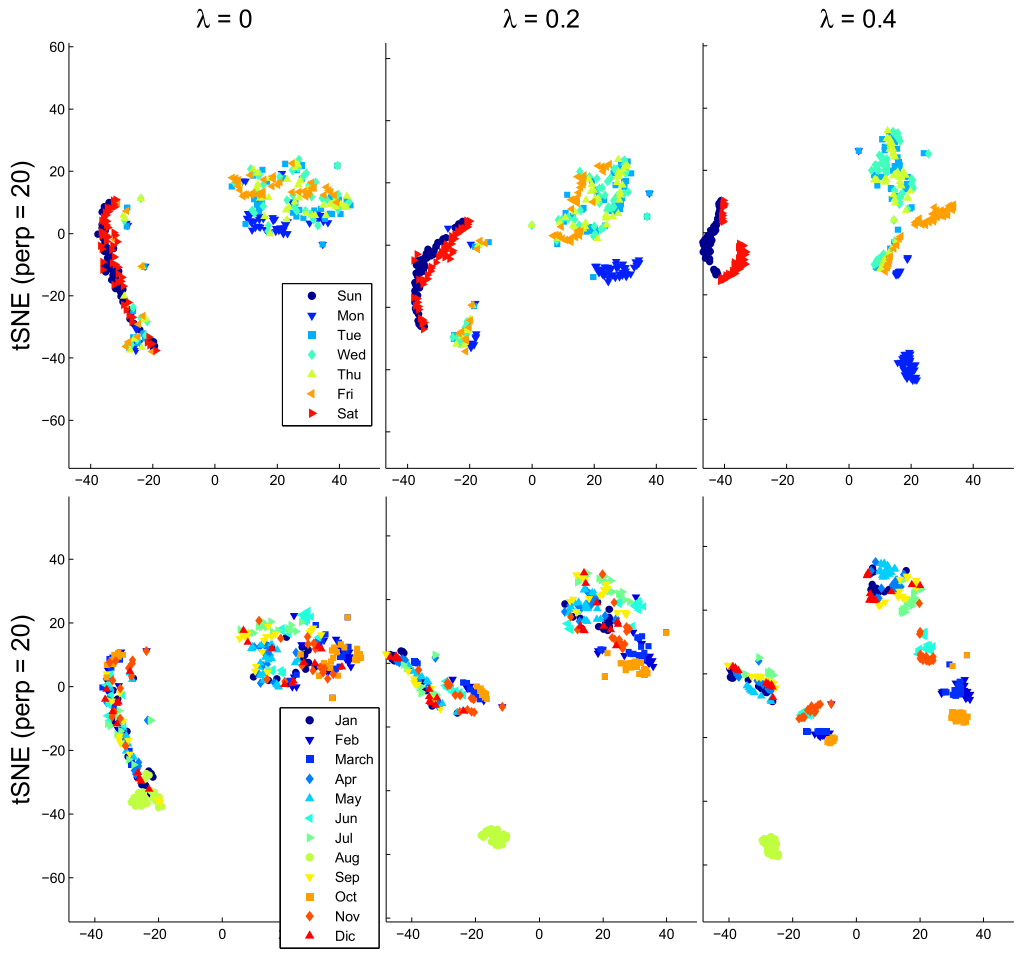

Figure 7: *t*-SNE projections for *eCons* dataset with a weighted extension applied for several values of λ using weekday (top) and month (bottom) labels.

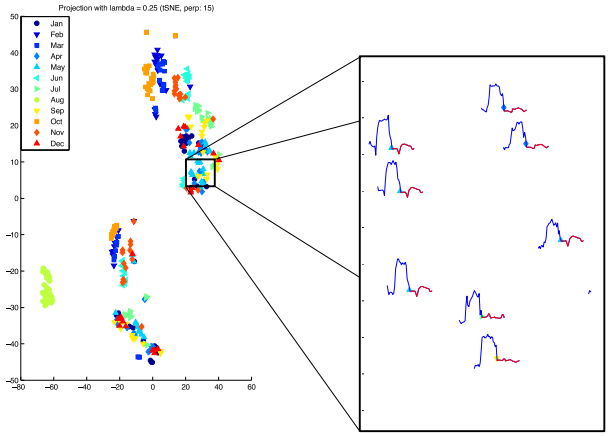

Figure 8: Projection with extended data by month (left) and zoomed representation of the projection (right), showing the features of each item as a sparkline, original (blue) and extended (red).

respect to the original.

Fig. 8 shows a part of a similar projection where the values of the features are plotted as a *sparkline* over each item, the original values (in blue) and the extended values (in red). This allows an easy comparison between similarities of the projected points. Points inside a class are topologically organized by intra-class similarities, which are given by the original features (blue). The inter-class organization between classes varies depending on the extended values (red) according to the value of λ, whose highest value (set to 1) corresponds to the pure projection of the centroids of the classes.

4.3. Extension of selected features

In this experiment we investigated the effect of simple extensions based on selected features using the *yeast* dataset [40] which is commonly used by ML and the visualization community as a benchmark dataset. The main task is to predict the localization site of proteins. Given the class labels, one thing the analyst can do is to study the distribution of the data values over different dimensions and detect discriminative features. This can often be achieved by examining visual representations of the distributions such

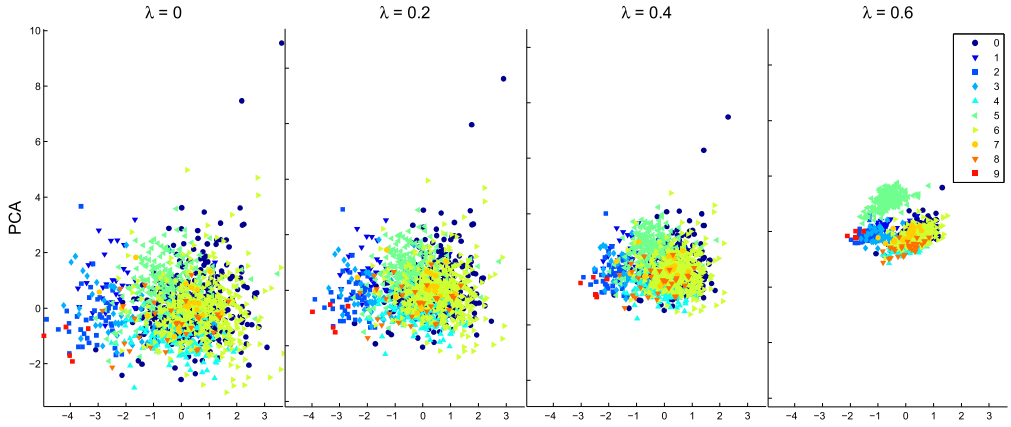

Figure 9: PCA projections for *yeast* dataset with a weighted extension of feature 4 for several values of λ using classes information.

as box plots or parallel coordinates. Furthermore other strategies of feature selection [41] can be used in cases where a multidimensional visualization can not be performed, such as relevance learning [35] or even domain knowledge of the user. In the current example, it is observed that objects in class 5 tend to have high values in dimension 4. This leads to our next experiment to extend only one feature –class mean of dimension 4– to see if the extended feature space leads to better visual quality.

Fig. 9 shows the resulting projections of the extended feature space using PCA. The left figure of $\lambda = 0$ is the projection of the original data. The rest projections are based on the weighted extension of mean values of dimension 4 over different classes. As one can see, overall the projection for $\lambda = 0.6$ is less cluttered. In particular, class 5 (in green) is much better separated from the rest of the classes.

The selection of a suitable λ value is made not only using a visual interpretation of the projection by the user, but also its evaluation performed by quality measures. The selection of this parameter takes into account the visual improvements of the projection whilst generally preserving the structure. In Fig. 10 the evolution of the measures used here can be seen for different values of λ. As the average of k-ary measure (Q_{avg}) equal to 1 means a perfect embedding, $1 - Q_{avg}$ is taken in order to show all lines with similar trends, i. e. the lower the value the better. It can be seen, with the evolution of λ, a remarkable improvement of visual measures (overlap area and density) and slightly worse structural measures (stress and k-ary). In

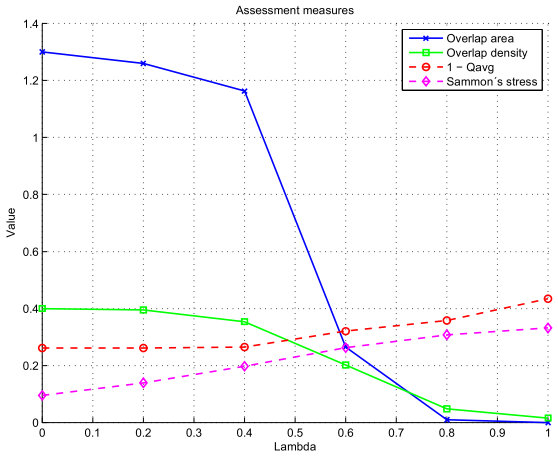

Figure 10: Quality measures of PCA projections of *yeast* data using weighted extension of the feature 4 for several values of λ.

this case, the analyst may want to select $\lambda = 0.6$ because it is the lowest λ parameter value that provides visual enhancements with small variations for the rest of the measures.

4.4. Supervised DR methods

In the last experiment we applied two supervised DR techniques, *NCA* and *MCML*, on some of the selected datasets. The same class information is used for the projection both the original data and data with weighted extensions.

3D clusters example. The method is applied to 3D clusters example from Section 4.1.1, the resulting projections with several values of λ are displayed in Fig. 11 for both techniques *NCA* (first from the top) and *MCML* (second). The regularization parameter of *NCA* is set to 0. As it can be seen in the figure, the weighted extension emphasize the class separation using both DR techniques in a similar way.

synthetic-gaussian dataset. Fig. 11 (third from the top) shows the projections generated using *MCML*. As one can see, even with a supervised DR method, the original data projection can be still quite cluttered ($\lambda = 0$). By

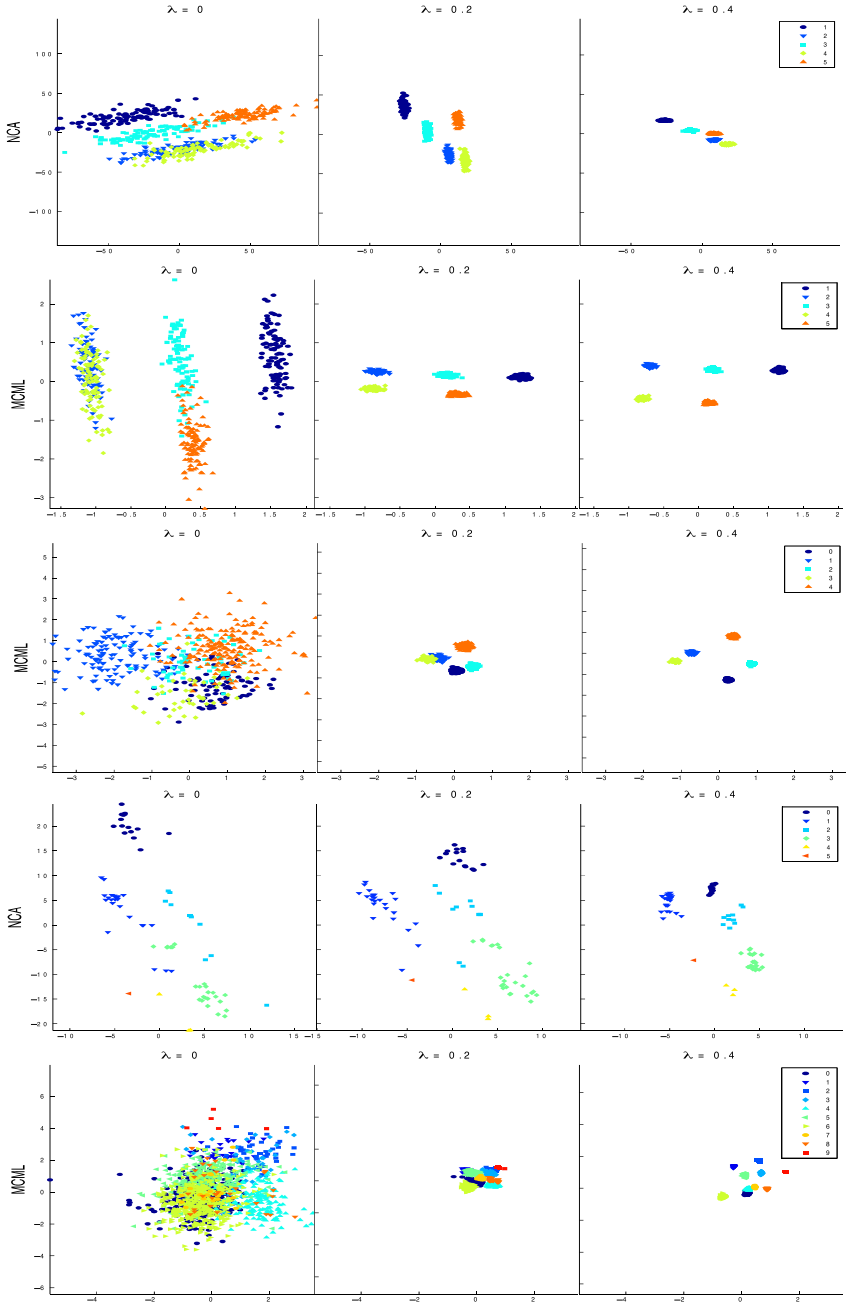

Figure 11: Projections with weighted extension for several λ values using labels. For *3D clusters* using NCA (first from the top) and MCML (second); for *synthetic-gaussian* using MCML (third); for *hiv* using NCA (fourth); and for *yeast* dataset using MCML (bottom).

transforming the original feature space with weighted extension (in this case, $\lambda = 0.4$), the group visual quality can be improved substantially.

hiv. The hiv dataset, which was used in [31], describes socio-economic properties of countries that are classified into HIV risk groups. The data has 159 attributes and contains objects that belong to 6 different classes. We project the data using *NCA*, in this case its regularization parameter is set to 0.001. The resulting projections are shown in the fourth projections of the Fig. 11. Again, although the groups are well-separated in the original projection, the projection with $\lambda = 0.2$ enhances the inter-group separation.

yeast. Fig. 11 (bottom) shows the projection of this dataset using *MCML*. While the projection based on the original dataset is rather crowded and one can hardly see any patterns, the projection of extended feature space ($\lambda = 0.4$) provides a much clearer view of the grouping information in the data.

4.5. Evaluation of embeddings

The performance of the projections is evaluated by arithmetic measures, described in Section 2.2. In this paper we select four measures, including the *Sammon's stress* [24] and *k-ary* [27] measure for assessing the structural preservation, and the *overlapping density* and *overlapping area* measures [8] for assessing the visual clarification. Although other approaches that represent the structural preservation and distortions could also be used for analyzing the quality of the result projections [42, 43].

The measures were computed for projections obtained using several values of λ. They are represented in line charts, similar to Fig. 10, where lower values imply improvements in the measures. Figures 12 to 15 graphically show the measures for *synthetic-gaussian*, *hiv*, *eCons (Months)*, and *yeast* datasets, respectively. The measures for *PCA* and *t-SNE* techniques are shown at the top and similarly for *NCA* and *MCML* methods at the bottom. Out of the range values were scaled in order to an effective comparison.

The result shows that in general extending feature space using class related statistical values leads to better visual quality in the final projection. Less overlapping points (reduced overlapping density measure), and group boundaries are overlapped less (reduced overlapping area measure). The data structure is less well-maintained in the new projection, especially the global pairwise distance (as indicated by the Sammon's stress measure). On

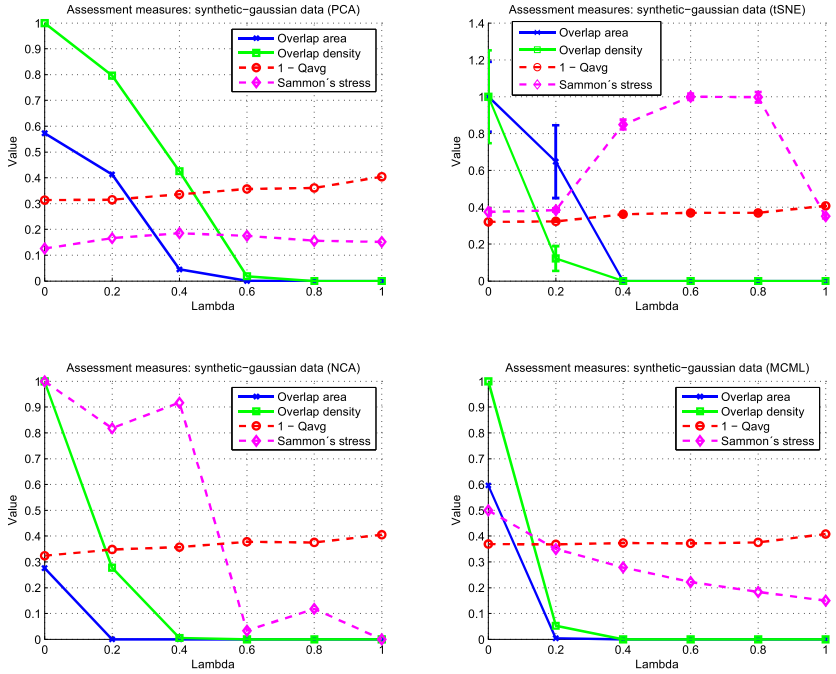

Figure 12: Assessment measures using PCA, t-SNE (top), NCA, and MCML methods (bottom) for *synthetic-gaussian* dataset

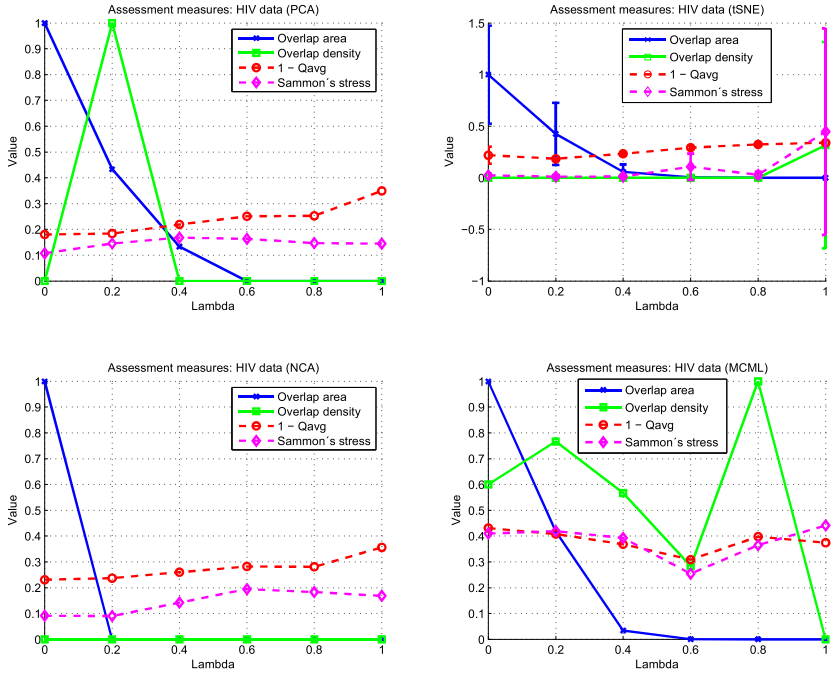

Figure 13: Assessment measures using PCA, t-SNE (top), NCA, and MCML methods (bottom) for *hiv* dataset

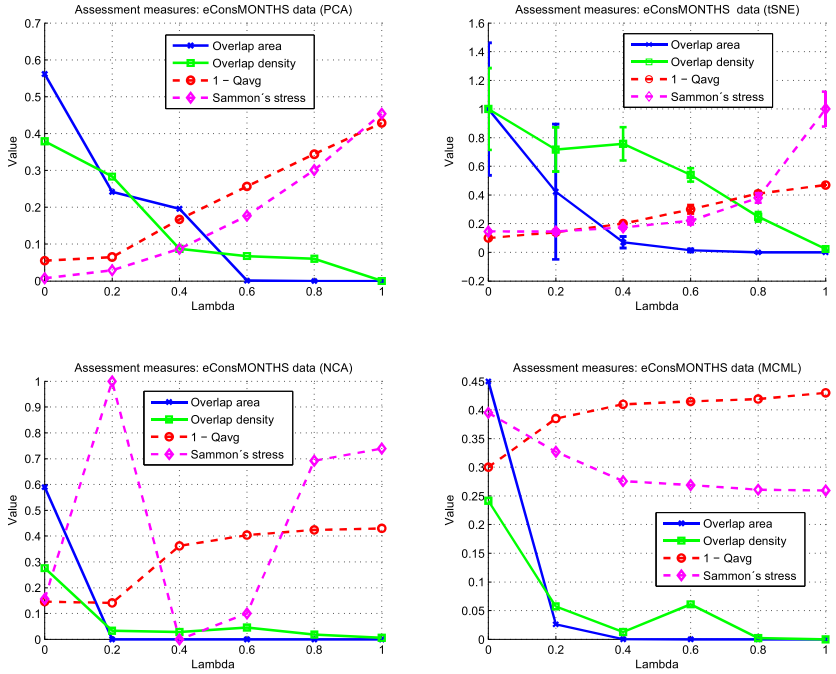

Figure 14: Assessment measures using PCA, t-SNE (top), NCA, and MCML methods (bottom) for *eCons (Months)* dataset

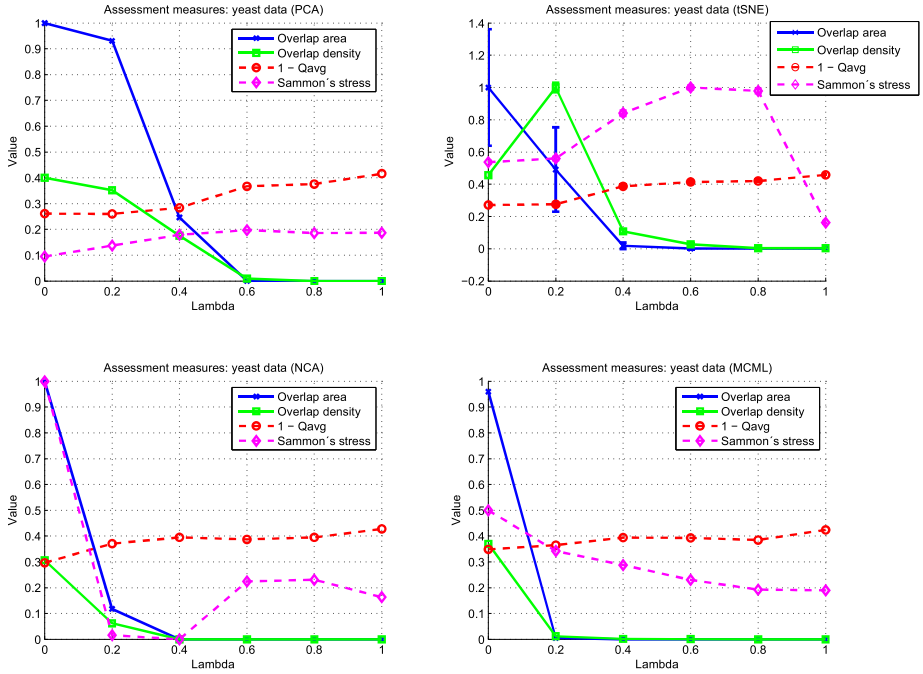

Figure 15: Assessment measures using PCA, *t*-SNE (top), NCA, and MCML methods (bottom) for *yeast* dataset

the other hand, in many cases the k-ary measures stay reasonably unchanged after transformation which means overall structural preservation in terms of both global distance and local neighborhood. The quality evaluation helps the analyst better understand the distortion caused by the transformation, and evaluate the quality gain in terms of visual clarification. A trade-off can be made fairly easily if a similar "quality graph" is provided during the analysis process.

Besides, the λ value itself gives a good indication of the "degree of distortion". By modifying the λ parameter the user can gradually control the extension or come back any previous point. This allows one to track the variations in the projections by smooth transitions, and to be aware of the trade-off between the original structure preservation and visual quality.

In addition to evaluate the projections with the quality graph explained before, there are more approaches that can be used to visualize the quality in the projection. For instance, the evaluation of a Self-Organizing Map can be visualized employing the U-Matrix [44]. This idea provides information to the user about the underlying structure preservation with respect to the original data into the embedding. In a similar way, a point-wise quality evaluation is proposed in [42] using a rank-based criteria that allows to highlight erroneous regions in the visualization. Using this approach, a mean error can be computed for each point and encoded as color in the projection. In Fig. 16 this evaluation of the quality is shown for the *eCons* data example with the labels of weekday where some points reveal worse structural quality with the appliance of the method. This useful visualization helps the user to be aware of the errors so that the control parameter can be set in a final value easily with a direct evaluation of the projection.

5. Discussion

Given an initial projection, the proposed approach allows the user to generate new projections with improved visual quality by integrating new group information into the DR process, assuming the new information is validated by the user and provides knowledge related to the analyzed tasks. The method can be applied when the grouping information can not be fully revealed by the distance measures that are used to compute the projection due to noise and irrelevant information. Another advantage of the proposed method is the preservation of local structure. As the distances between points within to the same group are not altered by the transformation, new projec-

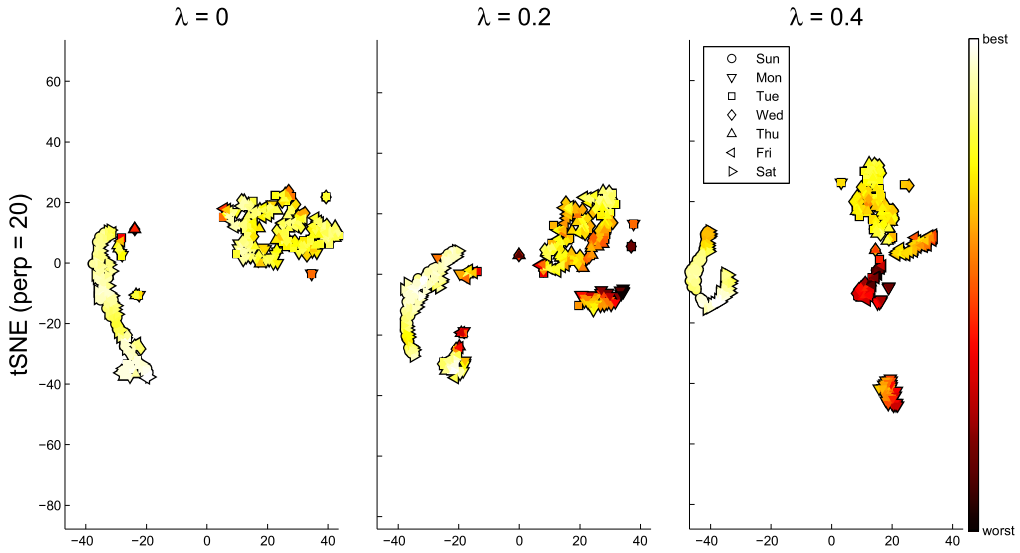

Figure 16: Point-wise quality visualization for the evaluation of the example with *eCons (weekday)* dataset for several λ values.

tions based on transformed data preserve the local structure in the original data. In addition, the DR assessment using numerical quality measures (see Section 4.5) gives an idea of the structural variations with respect to the original data and the visual improvements produced by the method.

From the visualization point of view, one may argue that given grouping information it is easier to use color or shape to differentiate groups in the projection. However, the color-coding and shape-coding approach do not solve the cluttering problem. When points are over-plotted in a visual display, the readability of the projection is still not much improved. Our approach helps to reduce cluttering in the projection and provides a more efficient visual channel for assessing relative distances between objects and classes. Furthermore, using space to separate classes makes it possible to apply other interaction mechanisms such as hovering over points to get contextual information or area selection to calculate aggregated values.

The transformation process is controllable via a weighting parameter λ. When $\lambda = 0$, the projection is purely based on the original data. As λ value enlarges, the method gradually increases the influence of the grouping information. When λ value reaches 1, only class information is used for computing projection hence all points collapse to their corresponding centroids. The an-

alyst can start with an initial projection ($\lambda = 0$) and gradually increase the λ value in order to achieve clearer group separation in the projection space. Such a process can be easily facilitated by an interactive graphical interface with a sliding bar.

The role of interaction is a key aspect within this context. The interactive changes of λ values allow the user to go back and forth as much as needed. This reversible process helps to keep in mind the initial reference view at each stage. In addition, the smooth variations of the points allow a continuous object tracking that can be performed with animation improving its graphical perception [45]. The user can examine the original data structure ($\lambda = 0$), the projection of class centroids ($\lambda = 1$), and intermediate views ($0 < \lambda < 1$) that allow to get new insights not available with a single DR tool. This not only can be used to achieve an interactive grouping separation but also to understand which part of the overlap of the classes is produced by the projection or is an actual characteristic of the multidimensional data. We further note that appropriate methods for visualization of projection qualities (e.g. based on projection stress), have been developed [44, 46, 47, 42, 43] and can be combined with our interactive approach. Especially in combination with interactive setting of λ values, a dynamic visualization of projection quality will help the analyst to assess the distortion introduced and strike a balance between data distortion and a de-cluttered projection.

For the interactive approach, it is always desirable to have smooth transition between views when the λ value is updated. This can be challenging due to computational time required to generate new projections, especially when the data is large in size and/or dimensionality. For example, PCA has a complexity of $O(d^3)$ where d is the number of dimensions, so when the dimensionality of data is very high, some preprocessing stage such as feature selection may be required to reduce the dimensionality of the data. Another example is the t-SNE approach, the original t-SNE algorithm has a complexity of $O(n^2)$ where n is the number of objects in the data. Although some recent work reduced the complexity of t-SNE to $O(n \log n)$ [34, 48], the method can still fail to support smooth transitions when n is large. In such case, sampling may be needed prior to the computation to reduce the computational load. Another possible solution to improve the scalability of the proposed approach is to pre-compute a series of projections with increasing λ values.

6. Conclusions

In this paper we propose a simple but effective approach that supports projection-based data analysis. The proposed interactive analysis framework extends traditional dimensionality reduction approaches that transform multi-dimensional data to a lower-dimensional visual display as a static view to an interactive visual display that allows the analyst to gradually modify the projection by incorporating grouping information. The proposed approach differs from traditional supervised DR methods in such a way that the user has more control over the analysis process. For example, they perform an extension of the features based on classes information and adjust the weight between original and extended feature space before projection so that the influence of class knowledge can be changed in the final projection. To bring more transparency to the analysis, the framework also integrates various quantitative measures to help analysts judge the quality of generated projection both in terms of structural preservation and visual clarification.

A number of experiments were carried out to evaluate the effectiveness of the proposed approach, covering different types of datasets, both supervised and unsupervised DR techniques, under different weighting conditions, and under different use case scenarios. The resulting projections are evaluated both visually and using quantitative measures that compute the structural preservation and visual quality. The experimental results indicate that the proposed methods not only lead to improved visual quality but also preserve the local neighborhood reasonably well. The resulting projections show the incorporation of meaningful information in a transparent manner. This provides efficiency in the visual analytics process for pattern recognition, fast identification of class labels and a better understanding of the data.

Future work includes exploring more interactive visualization techniques, the design of more sophisticated extension strategies that are tailor-made to the nature of data for improving the effectiveness of the methods, and to develop a wider range of quality measures for evaluating the projections. Moreover, a user study is planed to ascertain the usability of the proposed technique.

References

[1] J. Lee, M. Verleysen, Nonlinear dimensionality reduction, Springer, 2007.

[2] L. Van der Maaten, An introduction to dimensionality reduction using matlab, Report 1201 (07-07) (2007) 62.

[3] D. L. Donoho, High-dimensional data analysis: the curses and blessings of dimensionality, in: Proceddings of American Mathematical Society Conf. Math Challenges of the 21st Century (2000), 2000.

[4] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (7) (1936) 179–188.

[5] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighbourhood components analysis, Advances in Neural Information Processing Systems 17.

[6] A. Globerson, S. Roweis, Metric learning by collapsing classes, Advances in neural information processing systems 18 (2006) 451.

[7] D. A. Keim, J. Kohlhammer, G. Ellis, F. Mansmann, Mastering The Information Age - Solving Problems with Visual Analytics, Eurographics, 2010.

[8] M. Schaefer, L. Zhang, T. Schreck, A. Tatu, J. A. Lee, M. Verleysen, D. A. Keim, Improving projection-based data analysis by feature space transformations, in: Proceedings of the SPIE Visualization and Data Analysis (VDA), 2013.

[9] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, R. Chang, iPCA: an interactive system for PCA-based visual analytics, Computer Graphics Forum 28 (3) (2009) 767–774.

[10] J. Choo, H. Lee, J. Kihm, H. Park, ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction, in: Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on, 2010, pp. 27 –34.

[11] P. Joia, F. Paulovich, D. Coimbra, J. Cuminato, L. Nonato, Local affine multidimensional projection, Visualization and Computer Graphics, IEEE Transactions on 17 (12) (2011) 2563–2571.

[12] D. Perez, L. Zhang, M. Schaefer, T. Schreck, D. A. Keim, I. Diaz, Interactive Visualization and Feature Transformation for Multidimensional

Data Projection, in: Proc. EuroVis Workshop on Visual Analytics Using Multidimensional Projections, 2013.

[13] A. Endert, C. Han, D. Maiti, L. House, S. Leman, C. North, Observation-level interaction with statistical models for visual analytics, in: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, IEEE, 2011, pp. 121–130.

[14] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, R. Minghim, Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data, Computer Graphics Forum 31 (3pt4) (2012) 1345–1354.

[15] F. Paulovich, C. Silva, L. Nonato, User-centered multidimensional projection techniques, Computing in Science Engineering 14 (4) (2012) 74–81.

[16] S. Johansson, J. Johansson, Interactive dimensionality reduction through user-defined combinations of quality metrics, Visualization and Computer Graphics, IEEE Transactions on 15 (6) (2009) 993–1000.

[17] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, T. Möller, Dimstiller: Workflows for dimensional analysis and reduction, in: Proc. IEEE Conf. Visual Analytics Science and Technology (VAST), Vol. 1, Citeseer, 2010.

[18] G. M. Mamani, F. M. Fatore, L. G. Nonato, F. V. Paulovich, User-driven feature space transformation, in: Computer Graphics Forum, Vol. 32, Wiley Online Library, 2013, pp. 291–299.

[19] E. Brown, J. Liu, C. Brodley, R. Chang, Dis-function: Learning distance functions interactively, in: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, 2012, pp. 83–92.

[20] S. Bremm, T. von Landesberger, J. Bernard, T. Schreck, Assisted descriptor selection based on visual comparative data analysis, in: Computer Graphics Forum, Vol. 30, Wiley Online Library, 2011, pp. 891–900.

[21] S. Bremm, T. von Landesberger, M. Heß, T. Schreck, P. Weil, K. Hamacherk, Interactive visual comparison of multiple trees, in: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, IEEE, 2011, pp. 31–40.

[22] E. Bertini, A. Tatu, D. Keim, Quality metrics in high-dimensional data visualization: An overview and systematization, Proceedings of the IEEE Symposium on IEEE Information Visualization (InfoVis) 17 (2011) 2203–2212.

[23] J. A. Lee, M. Verleysen, Quality assessment of dimensionality reduction: Rank-based criteria, Neurocomputing 72 (7) (2009) 1431–1443.

[24] J. W. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput. 18 (5) (1969) 401–409.

[25] J. Kruskal, Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation', in: R. Milton, J. Nelder (Eds.), Statistical Computation, Academic Press, New York, 1969, pp. 427–440.

[26] J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: An experimental study, in: G. Dorffner, H. Bischof, K. Hornik (Eds.), Proceedings of ICANN 2001, Springer, Berlin, 2001, pp. 485–491.

[27] J. Lee, M. Verleysen, Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods, in: Y. Saeys, H. Liu, I. Inza, L. Wehenkel, Y. Van de Peer (Eds.), JMLR Workshop and Conference Proceedings (New challenges for feature selection in data mining and knowledge discovery), Vol. 4, 2008, pp. 21–35.

[28] L. Saul, S. Roweis, Think globally, fit locally: Unsupervised learning of nonlinear manifolds, Journal of Machine Learning Research 4 (2003) 119–155.

[29] J. Venna, S. Kaski, Nonlinear dimensionality reduction as information retrieval, in: M. Meila, X. Shen (Eds.), Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007), Omnipress, San Juan, Puerto Rico, 2007, pp. 568–575.

[30] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, D. Keim, Combining automated analysis and visualization techniques for effective exploration of high-dimensional data, in: Proc. IEEE Symp. Visual Analytics Science and Technology (VAST), 2009, pp. 59–66.

[31] M. Sips, B. Neubert, J. P. Lewis, P. Hanrahan, Selecting good views of high-dimensional data using class consistency., Comput. Graph. Forum 28 (3) (2009) 831–838.

[32] A. Moreira, M. Y. Santos, Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points, in: GRAPP 2007 : proceedings of the International Conference on Computer Graphics Theory and Applications. INSTICC Press, 2007. ISBN 978-972-8865-71-9., 2007, pp. 61–68.

[33] M. Li, J. T. Kwok, B.-L. Lu, Making large-scale nystrom approximation possible., in: J. Furnkranz, T. Joachims (Eds.), ICML, Omnipress, 2010, pp. 631–638.

[34] Z. Yang, J. Peltonen, S. Kaski, Scalable optimization of neighbor embedding for visualization, in: ICML (2), 2013, pp. 127–135.

[35] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, M. Biehl, Limited rank matrix learning, discriminative dimension reduction and visualization, Neural Networks 26 (2012) 159–173.

[36] D. G. Kendall, A survey of the statistical theory of shape, Statistical Science 4 (2) (1989) 87–99.

[37] M. Sedlmair, A. Tatu, T. Munzner, M. Tory, A taxonomy of visual cluster separation factors, in: Computer Graphics Forum, Vol. 31, Wiley Online Library, 2012, pp. 1335–1344.

[38] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

[39] J. J. Van Wijk, E. R. Van Selow, Cluster and calendar based visualization of time series data, in: Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on, IEEE, 1999, pp. 4–9.

[40] C. Blake, C. J. Merz, {UCI} repository of machine learning databases.

[41] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research 3 (2003) 1157–1182.

[42] B. Mokbel, W. Lueks, A. Gisbrecht, B. Hammer, Visualizing the quality of dimensionality reduction, Neurocomputing 112 (2013) 109–123.

[43] M. Aupetit, Visualizing distortions and recovering topology in continuous projection techniques, Neurocomputing 70 (7) (2007) 1304–1330.

[44] A. Ultsch, H. P. Siemon, Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis, in: INNC Paris 90, Universitat Dortmund, 1990, pp. 305–308.

[45] J. Heer, G. G. Robertson, Animated transitions in statistical data graphics, Visualization and Computer Graphics, IEEE Transactions on 13 (6) (2007) 1240–1247.

[46] S. Lespinats, M. Aupetit, Checkviz: Sanity check and topological clues for linear and non-linear mappings, Comput. Graph. Forum 30 (1) (2011) 113–125.

[47] T. Schreck, T. von Landesberger, S. Bremm, Techniques for precision-based visual analysis of projected data, Information Visualization 9 (3) (2010) 181–193.

[48] L. van der Maaten, Barnes-hut-sne, arXiv preprint arXiv:1301.3342.

*Biography of the author(s)
Click here to download Biography of the author(s): vitaes.docx

Daniel Pérez is currently a PhD-student at University of Oviedo, Spain, in the research group of supervision and diagnos c of industrial processes. He studied industrial engineering at University of Oviedo, where he received his M.Eng in 2007, later he became a research assistant in Electrical Engineering Department at the University of Oviedo. His main research interests are informa on visualiza on, machine learning, and visual analy cs. He joined in his current group in 2011 and is currently working in visualiza on of high-dimensional data projec ons.

Dr. Leishi Zhang is a Lecturer in Visual Analy cs at Middlesex University, UK. She received her PhD in Computer Science from Brunel University, UK for her work in me series data analysis and visualiza on. Her research interests include me series data modelling, high-dimensional data projec on and interac ve visualiza on of subspace clusters. She has worked on various research and industrial projects in visual analy cs and published her research in a number of interna onal journals and conferences.

Ma hias Schäfer is a PhD-student at University of Konstanz in the Visualiza on and Data Analysis Group. He is working at Department of Computer and Informa on Science and his current research interest are informa on visualiza on, data mining, mul media- and mul dimensional-databases and visual analy cs.

Dr. Tobias Schreck is an assistant professor of visual analy cs in the Department of Computer and Informa on Science at the University of Konstanz, Germany. His research interests include visual search and analysis in me-oriented, high-dimensional, and 3D object data, with applica ons in data analysis and mul media retrieval. Schreck received a PhD in computer science from the University of Konstanz. He is a member of IEEE.

Daniel A. Keim is a full professor in the Department of Computer Science at the University of Konstanz, Germany and chair of the university's Visualiza on and Data Analysis Group. His research interests include visual analy cs, informa on visualiza on, and data mining. Keim received a PhD in computer science from the University of Munich, Germany. He is a member of the IEEE Computer Society and a coordinator of the German strategic research ini a ve on scalable visual analy cs.

Ignacio Díaz is associate professor of the Electrical Engineering Department at the University of Oviedo since 2004. He received a M.Eng in 1995 and his Ph.D in industrial engineering in 2000 by the University of Oviedo. His main research interests are the applica on of data visualiza on and intelligent data analysis algorithms to industrial problems. He has led several R&D projects nanced by the Spanish government and the European Union and published his research in indexed journals, as well as numerous publica ons in interna onal conferences. Professor Díaz is member of the IEEE since 1997.

*Photo of the author(s)

Click here to download high resolution image

*Photo of the author(s)
Click here to download high resolution image

*Photo of the author(s)
Click here to download high resolution image

*Photo of the author(s)
Click here to download high resolution image

*Photo of the author(s)
Click here to download high resolution image

*Photo of the author(s)

Click here to download high resolution image

Supplementary Material for on-line publication only

Click here to download Supplementary Material for on-line publication only: video.mov

LaTeX Souce Files

Click here to download LaTeX Souce Files: NC-R3-SUBMISSION.zip